Assessment of phylogenetic structure in genome size - gene content correlations

Prasad, V R ; Isler, K

Abstract: Gene content and gene-coding percentage can be predicted from genome size in newly sequenced organisms. Here, we investigate whether these predictions are influenced by phylogenetic relationships between the involved species. Combining a highly resolved phylogenetic tree with a large compilation of gene content data, our results reveal the presence of significant phylogenetic structure in the correlations between genome size and gene content in both bacteria and eukaryotes. The variation in log(gene content) explained by log(genome size) in combination with phylogeny was found to be 97% in bacteria and 55% in eukaryotes. Further, in bacteria, gene-coding percentages are only significantly correlated to genome size if phylogenetic information is taken into account in the analyses. These findings support the usage of phylogenetic correlation models for gene content predictions.

DOI: https://doi.org/10.1139/g2012-019

Assessment of phylogenetic structure in genome size-gene content correlations

NOTE

Vibhu Ranjan Prasad[1] and Karin Isler[2]

[1] Institute of Molecular Life Sciences, University of Zurich-Irchel, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[2] Anthropological Institute and Museum, University of Zurich-Irchel, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Published in: Genome (2012) 55, p. 391-395

Corresponding author:

Vibhu Ranjan Prasad, Institute of Molecular Life Sciences, University of Zurich-Irchel, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

phone/fax: +41 (0)44 63 55396

email: vibhuranjan.prasad@uzh.ch

Keywords: genome size, gene content, gene-coding percentage, phylogeny, independent contrasts

Running head: Phylogeny predicts genome size-gene content correlation

List of abbreviations:

IC: independent contrasts

PGLS: phylogenetic least-squares regression

Karin
Delet
Figure
1.3 pag

ABSTRACT

Gene content and gene-coding percentages can be predicted from genome size in newly sequenced organisms. Here, we investigate whether these predictions are influenced by phylogenetic relationships between the involved species. Combining a highly resolved phylogenetic tree with a large compilation of gene content data, our results reveal the presence of significant phylogenetic structure in the correlations between genome size and gene content in both bacteria and eukaryotes. The variation in log-gene content explained by log-genome size in combination with phylogeny was found to be 97% in bacteria and 55% in eukaryotes. Further, in bacteria gene-coding percentages are only significantly correlated to genome size if phylogenetic information is taken into account in the analyses. These findings support the usage of phylogenetic correlation models for gene content predictions.

TEXT

Gene content, the number of genes coding for proteins, is correlated with genome size in both non-eukaryotes and eukaryotes (Lynch and Conery 2003; Konstantinidis and Tiedje 2004; Gregory 2005). More detailed knowledge about this relationship would be important, as it is highly desirable to be able to predict gene content in species with massive genome size. Recently, Hou and Lin (2009) proposed a linear relationship between log(genome size) and log(gene content) for non-eukaryotes and a steeper logarithmic relationship for eukaryotes, both with high correlation of R^2>0.91. Similar relationships were found between genome size and the total number of genes, or, inversely, with the amount of non-coding DNA. Their compilation includes 55 eukaryote and 1055 non-eukaryote species with completely sequenced and annotated genomes. However, given that these species are phylogenetically related to various degrees, phylogenetic information should be included in the correlation analyses, which has not been attempted yet. Here, we examine the correlations between genome size and gene content or coding percentage of genome, applying methods for "phylogenetically informed analysis" that have been developed for correlative studies in other fields (e.g. Felsenstein 1985; Garland et al. 2005). The method of phylogenetic generalized least-squares (PGLS) is currently the most frequently used approach for phylogenetic regression (Nunn 2011), as it allows the assumed model of evolution to deviate from Brownian motion (Pagel 1999; Freckleton et al. 2002). To test for phylogenetic structure in the data, PGLS simultaneously estimates the parameter λ for a given combination of data and tree with a maximum likelihood approach. A λ value significantly different from 0 indicates that the data values cluster according to the structure of the given phylogenetic tree, and thus that phylogenetic regression is warranted. If λ is close to 1, PGLS yields roughly the same results as the classic independent contrasts (IC) method (Freckleton et al. 2002).

Phylogenetic and non-phylogenetic correlations between genome size and both gene-content and gene-coding percentages are shown in Table 1 and illustrated in Figure 2. For the correlation between gene-content and genome size, λ values are close to 1 and significantly different from 0 in both eukaryotes and non-eukaryotes, indicating that a phylogenetic approach is needed for this analysis. Thus, results of phylogenetic least-squares regression (PGLS) are almost identical to independent contrasts (IC) results in all three groups. Overall, the correlation coefficients and p-values are very similar in phylogenetic vs. non-phylogenetic analyses, although for eukaryotes, our analysis yields a weaker correlation in phylogenetic regression (R^2=0.55, p=.0005) than in raw species regression (R^2=0.84, p<.0001). The variation in log-gene content explained by log-genome size in combination with phylogeny was found to be 97% in bacteria and 55% in eukaryotes.

For the correlation of gene coding percentage with genome size, λ values are close to 1 and differ significantly from zero for the bacterial and the combined dataset, but not for eukaryotes (Table 1). Thus, PGLS results are more similar to raw correlations in the latter group. Interestingly, the very low correlation between genome size and gene-coding percentage in bacteria is only significantly different from zero if phylogenetic relationships are taken into account. Log-genome size predicts 81% of the variation in log-gene-coding percentage in eukaryotes, but only about 7% in bacteria (for prediction equations, see Supplementary Information).

Our results indicate the presence of phylogenetic structure in the correlations of genome size with gene content in both bacteria and eukaryotes, and in the correlation of genome size with gene-coding percentage in bacteria. The absence of phylogenetic structure in eukaryote gene-coding percentages may be due to the relatively small number of species in our sample (n=19), which are all very distantly related (e.g. 1 bird, 1 fish, 1 worm, 1 algae and 4 mammals). In bacteria, the sample is larger (n=82) and the variation in the degree of

relationship between the included species is higher, yielding a better estimate of phylogenetic structure. In contrast to earlier studies (Hou and Lin 2009), we found that the gene-coding percentage is also significantly correlated to genome size in bacteria, but only if phylogenetic relationships between species are taken into account in the analysis.

We therefore conclude that phylogenetic structure should be considered in any attempt to predict gene content or gene-coding percentage from genome size in organisms. Although the present sample is rather small, we expect that the amount of phylogenetic structuring will be even larger if the sample is expanded by inclusion of more, closely related species. Additionally, considering phylogenetic relationships may provide insights into whether a particular deviation from the general trend is species-specific (e.g. due to recent polyploidy, cf. Otto 2007) or lineage-specific (e.g. an adaptive event in ancestral birds and bats, Hughes and Hughes 1995). Depending on the availability of larger datasets, future studies may disentangle different scenarios of phylogenetic inertia and adaptation in gene content - genome size variation across organisms, as has proven useful in comparataive studies of genome size (e.g. Oliver et al. 2007, Organ and Shedlock 2009).

Vibhu
Delet

METHODS

Taken from Hou and Lin (2009), the variables in the present study include genome size, protein coding gene number, and coding percentage. All data were log-transformed in order to increase normality of their distributions over a large range of values, following Hou and Lin (2009). Phylogenetic information was taken from the Interactive Tree of Life (Letunic and Bork 2007), a highly resolved phylogenetic tree based on 31 marker gene families which are found universally among 191 sequenced genomes. In this tree, the families which show multiple horizontal transfers and cause difficulty in alignment were removed and the

Karin
Delet

Karin
Delet
in the
transfo

phylogeny was constructed using Maximum Likelihood reconstruction as described in Ciccarelli et al. (2006). This approach was chosen as it allows to include a very diverse range of organisms in a consistent manner, although it may not yield a maximum sample size within lineages. Limited by the overlap between the two, the dataset on gene content and the phylogenetic tree, our sample comprises 101 organisms (82 bacteria and 19 eukaryotes, listed in the Supplementary Information). The pruned version of the tree used in this analysis is shown in Figure 1.

To test whether there is phylogenetic autocorrelation in the data, and thus whether a phylogenetic approach is warranted, the parameter λ was estimated using the CAIC package (Purvis and Rambaut 1995) in R 2.14 (R Development Core Team 2011). Varying from 0 to 1, λ close to 0 indicates that there is no phylogenetic signal in the data, whereas λ close to 1 indicates Brownian motion phylogenetic autocorrelation in the analysed traits. In other words, if the value of λ is found to be significantly different from zero, phylogenetic methods such as the here applied phylogenetic generalized least squares (PGLS) regression are warranted (Pagel 1999). PGLS fits a linear model to the data, which therefore must be transformed accordingly before analysis. Although Hou and Lin (2009) found a slightly better fit for a logarithmic relationship between log(gene cont) and log(genome size), in our smaller sample the difference between a logarithmic and a linear model was minimal. We therefore consistently used a linear model for all analyses. PGLS converts the phylogeny into a variance-covariance matrix, which is then included in the error term of the regression model. The resulting estimated regression parameters are "phylogenetically controlled" (Pagel 1999; Freckleton et al. 2002). For illustration, phylogenetically independent contrasts (IC; Felsenstein 1985; Garland et al. 1992) were calculated in Mesquite (Maddison and Maddison

2008) using the PDAP: PDTREE package (Midford et al. 2002) and analysed using the program JMP (JMP 2009).

Acknowledgements

Vibhu Ranjan Prasad would like to thank Dhirendra Mohan Prasad, Anubhuti Ranjan Prasad and Anand Mala Prasad for their continuous encouragement. We thank two anonymous reviewers for their helpful comments on an earlier version of the manuscript.

Literature cited

Ciccarelli, F.D., Doerks, T., Von Mering, C., Creevey, C.J., Snel, B., Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science, 311:1283-1287.

Felsenstein, J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1-15.

Freckleton, R.P., Harvey, P.H., Pagel, M. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. Am. Nat. 160:712-726.

Garland Jr, T., Bennett, A.F., Rezende, E.L. 2005. Phylogenetic approaches in comparative physiology. J. Exp. Biol. 208:3015-3035.

Garland, T., Harvey, P.H., Ives, A.R. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Syst. Biol. 41:18-32.

Gregory, T.R. 2002. A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the class Aves. Evolution, 56:121-130.

Gregory, T.R. 2005. Synergy between sequence and size in large-scale genomics. Nat. Rev. Genet. 6:699-708.

Hou, Y., Lin, S. 2009. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS One, 4:e6978.

Hughes, A.L., Hughes, M.K. 1995. Small genomes for better flyers. Nature 377:391.

JMP, 2009. Version 8.0. SAS Institute. Inc., Cary, NC.

Konstantinidis, K.T., Tiedje, J.M. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl. Acad. Sci. U.S.A. 101:3160-3165.

Letunic, I., Bork, P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23:127-128.

Lynch, M., Conery, J.S. 2003. The origins of genome complexity. Science, 302:1401-1404.

Maddison, W.P., Maddison, D.R. 2008. Mesquite: a modular system for evolutionary analysis. Evolution, 62:1103-1118.

Midford, P., Garland Jr, T., Maddison, W. 2002. PDAP: PDTREE package for Mesquite, version 1.00. Available from http://www. http://mesquiteproject.org/pdap_mesquite/ [accessed 21 July 2011].

Nunn, C.L. 2011. The Comparative Approach in Evolutionary Anthropology and Biology. University of Chcago Press, Chicago, IL.

Oliver, M.J., Petrov, D., Ackerly, D., Falkowski, P, Schofield, O.M. 2007. The mode and tempo of genome size evolution in eukaryotes. Genome Res. 17:594-601.

Organ, C.L, Shedlock, A.M. 2009. Paleogenomics of pterosaurs and the evolution of small genome size in flying vertebrates. Biol. Letters, 5:47-50.

Otto, S.P. 2007. The evolutionary consequences of polyploidy. Cell, 131:452-462.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature, 401:877-884.

Karin
Delet

Purvis, A., Rambaut, A. 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. CABIOS, 11:247-251.

R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Version 2.14.

Tables

Table 1. Correlation of log genome size with log gene content and log gene-coding percentage for eukaryotes, bacteria and all species using phylogenetic generalized least-squares (PGLS), a non-phylogenetic model (Raw) and independent contrasts (IC). Regression equations are shown in the Supplementary Information.

	Gene content					Gene-coding percentage				
	PGLS			Raw (λ=0)	IC (λ=1)	PGLS			Raw (λ=0)	IC (λ=1)
	λ	P-value of $\lambda \neq 0$	R^2 P-value	R^2 P-value	R^2 P-value	λ	P-value of $\lambda \neq 0$	R^2 P-value	R^2 P-value	R^2 P-value
Eukaryote (n=19)	0.97	0.03*	0.55 .0002*	0.82 <.0001*	0.51 0.0005*	0.001	1	0.819 <.0001*	0.819 <.0001*	0.46 0.003*
Bacteria (n=82)	0.84	0.0007*	0.97 <.0001*	0.97 <.0001*	0.96 <.0001*	0.999	<.0001*	0.068 0.017*	0.013 0.29	0.069 0.016*
All species (n=101)	0.97	<.0001*	0.74 <.0001*	0.798 <.0001*	0.725 <.0001*	0.967	<.0001*	0.264 <.0001*	0.77 <.0001*	0.08 0.003*

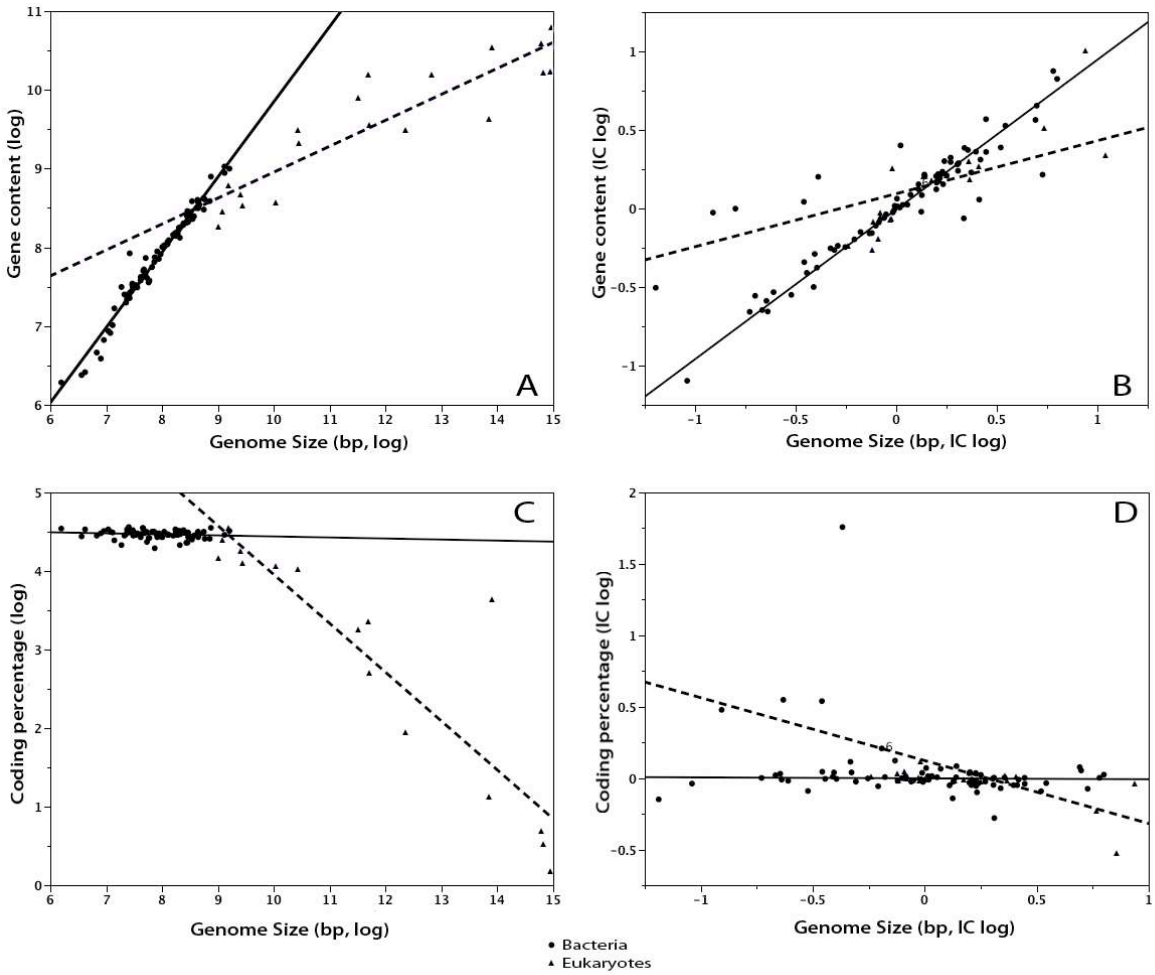
Figure legends

Fig 1. The highly resolved phylogenetic tree of 101 species used in this study for Independent Contrasts and PGLS.

Fig 2. Correlations of log (gene content) and log (gene-coding percentage) with genome size without phylogeny (A, C), and with phylogeny (B, D) using independent contrasts (IC). Dashed line is eukaryotic and solid line is bacterial fit. For statistics see Table 1.


PAUP 1++++
Arabidopsis thaliana
Oryza sativa
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Danio rerio
Gallus gallus
Pan troglodytes
Homo sapiens
Mus musculus
Rattus norvegicus
Dictyostelium discoideum
Thalassiosira pseudonana
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Eremothecium gossypii
Plasmodium falciparum
Cryptosporidium hominis
Giardia lamblia
Sulfolobus solfataricus
Aeropyrum pernix
Pyrobaculum aerophilum
Nanoarchaeum equitans
Methanosarcina mazei
Pyrococcus abyssi
Methanopyrus kandleri
Methanobacterium thermautotrophicum
Methanococcus maripaludis
Thermoplasma acidophilum
Helicobacter pylori 26695
Treponema pallidum
Bdellovibrio bacteriovorus
Mycoplasma gallisepticum
Ureaplasma parvum
Borrelia burgdorferi
Chlamydia trachomatis
Chlamydophila caviae
Chlamydophila pneumoniae TW183
Rhodopirellula baltica
Leptospira interrogans 56601
Bacteroides thetaiotaomicron
Porphyromonas gingivalis
Chlorobium tepidum
Corynebacterium diphtheriae
Mycobacterium bovis
Streptomyces avermitilis
Tropheryma whipplei TW0827
Bifidobacterium longum
Deinococcus radiodurans
Thermus thermophilus
Synechocystis sp PCC6803
Nostocsp PCC 7120
Synechococcus elongatus
Prochlorococcus marinus CCMP1378
Gloeobacter violaceus
Campylobacter jejuni
Wolinella succinogenes
Pseudomonas aeruginosa
Escherichia coli EDL933
Shigella flexneri 2a 2457T
Salmonella enterica
Photorhabdus luminescens
Haemophilus influenzae
Pasteurella multocida
Vibrio cholerae
Photobacterium profundum
Shewanella oneidensis
Xanthomonas campestris
Xanthomonas axonopodis
Xylella fastidiosa 9a5c
Ralstonia solanacearum
Bordetella pertussis
Neisseria meningitidis A
Chromobacterium violaceum
Nitrosomonas europaea
Coxiella burnetii
Blochmannia floridanus
Bradyrhizobium japonicum
Rhodopseudomonas palustris
Rhizobium loti
Brucella melitensis
Agrobacterium tumefaciens Cereon
Caulobacter crescentus
Rickettsia conorii
Wolbachia sp wMel
Desulfovibrio vulgaris
Geobacter sulfurreducens
Solibacter usitatus
Staphylococcus aureus Mu50
Bacillus anthracis
Oceanobacillus iheyensis
Listeria innocua
Streptococcus mutans
Streptococcus agalactiae V
Lactococcus lactis
Enterococcus faecalis
Clostridium acetobutylicum
Thermoanaerobacter tengcongensis
Fusobacterium nucleatum
Thermotoga maritima
Aquifex aeolicus

1


A
Gene content (log)
Genome Size (bp, log)
B
Gene content (IC log)
Genome Size (bp, IC log)
C
Coding percentage (log)
Genome Size (bp, log)
D
Coding percentage (IC log)
Genome Size (bp, IC log)
Bacteria
Eukaryotes

1